CLUSTERED KNOWLEDGE REPRESENTATION: INCREASING THE RELIABILITY OF COMPUTERIZED EXPERT SYSTEMS

by

Hong Yu

A thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Medical Informatics
The University of Utah

March 1990

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

SUPERVISORY COMMITTEE APPROVAL

of a thesis submitted by

Hong Yu

This thesis has been read by each member of the following supervisory committee and by majority vote has been found to be satisfactory.

[redacted] [redacted]

Peter J. H[redacted]

[redacted] [redacted]

Homer R. Warner

[redacted] [redacted]

[redacted] J. Lincoln

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

FINAL READING APPROVAL

To the Graduate Council of the University of Utah:

I have read the thesis of [redacted] Yu in its final form and have found that (1) its format, citations and bibliographic style are consistent and acceptable; (2) its illustrative materials including figures, tables and charts are in place; and (3) the final manuscript is satisfactory to the supervisory committee and is ready for submission to The Graduate School.

[redacted]

Peter J. [redacted]
Chair, Supervisory [redacted]

Approved for the Major Department

Homer R. Warner
Chair/Dean

Approved for the Graduate Council

B. Gale Dick
Dean of The Graduate School

ABSTRACT

I tested a group of frames intended for a medical diagnosis system called Iliad. This system is a microcomputer-based (Macintosh) medical expert system. The Iliad system contains a knowledge base, data dictionary, application programs and recent medical literature. Iliad is a Bayesian medical expert system. The system performs two functions for medical students: consultation and simulation.

Accuracy and reliability are major concerns for the development of a Bayesian expert system. The sequential Bayesian model is based on an assumption of conditional data independence. However, many disease findings are interrelated, and tend to co-occur. Some of these co-occurring findings describe pathophysiologic concepts, such as "lung consolidation." To handle these co-occuring findings, a new type of decision frames, called "clusters," have been included in the Iliad system. Clusters are rule-based decision frames which contain the conditional dependent findings. Clusters are used as findings in Bayesian frames, and thereby reduce the overconfidence that would result from including the conditional dependent findings directly. I hypothesized that the clusters would, in fact, significantly improve Iliad's diagnostic accuracy and reliability, compared to a non-clustered system. I tested this hypothesis by measuring the

reliability of pairs of clustered and nonclustered frames using real patient data.

The null hypothesis of my test assumed there was no difference between the clustered system of frames and the nonclustered system. This hypothesis was tested under two condition: The first condition used estimated probabilities for the frames. The second condition used actual probabilities measured from the data base. The test of both conditions allowed us to determine whether inaccurate statistical estimates might partly explain any unreliability or whether all unreliability was a result of conditional dependent findings. The test of frame reliability was developed by Hilden, et al.

According to my research, I found that the results generated by the clustered system were significantly more reliable than the results generated by the nonclustered system. Expert probability estimates were found to be inaccurate compared to actual measurements from the patient data. However, this inaccuracy did not explain the unreliability I found. This unreliability was due to conditional dependent findings. Some clustered frames remained unreliable on initial testing. When modified by reclusterring, these frames proved reliable. Reliability testing could be used during the knowledge engineering process to validate prototype frames.

To my new born baby.

TABLE OF CONTENTS

LIST OF TABLES

Number

LIST OF FIGURES

<u>Number</u>

ACKNOWLEDGMENTS

I wish to thank the members of my Supervisory Committee for their assistance. I would especially like to express my appreciation to Dr. Haug for his invaluable advice during my research work, Dr. Warner for his guidance and encouragement, and Dr. Lincoln for his invaluable help.

I would like to thank Dr. Charles Turner for his helpful suggestions and discussions.

My sincere thanks to my grandparents and parents who constantly trusted and supported me throughout my study.

Finally, I am deeply obliged to my husband, Chinli Fan, whose patience, support and encourgement were of great value to me.

CHAPTER 1

INTRODUCTION

Computerized medical diagnosis can be a useful tool in clinical and teaching activities. A large number of medical decision making systems have been developed in the past twenty years. Some of these medical decision making systems were designed with the capability to make decisions over a broad range of general internal medicine. Others were dedicated to performing specific tasks. In this document I describe one approach to testing the accuracy of these programs.

1.1 Overview of medical decision making systems

Most medical decision making systems are based on Bayesian probability calculations, optimum tree, linear logistic regression, rule-based as well as different combination and other methods. Examples of some of these types are discussed below.

Gorry and Barnett created a medical decision making system which not only accomplished diagnostic inference but also determined what tests could further clarify the diagnosis.[1] Because physicians rarely have enough initial information to make a satisfactory diagnosis, Gorry and Barnett devised a scheme to gather the additionally needed information. Bayes probability calculations were used to process the initial patient data.[2]

The MYCIN system was developed by Edward Shortliffe.[3] This system was the precursor to the present rule-based medical consultation (or decision) system, and this system was developed to advise physicians on antimicrobial therapies for patients with bacterial infections before the bacterial culture results were known. For this purpose, the knowledge base of MYCIN was comprised of therapeutic decision rules.

INTERNIST is an experimental computer-based diagnostic program which was developed at the University of Pittsburgh.[4] INTERNIST-I was designed to aid the physician when presented with the patient's initial history, results of a physical examination or laboratory findings. It could make multiple and complex diagnoses. The capabilities of the system derive from its extensive knowledge base and from heuristic computer programs that could construct and resolve differential diagnoses.

The building block for the INTERNIST-I knowledge base is the individual disease. For each diagnosis entered into the system, a disease profile is constructed. The disease profile consists of findings (symptoms, signs and laboratory abnormalities) that have been reported to occur in association with the disease. Two clinical variables are associated with each manifestation in an INTERNIST-I disease profile: an evoking strength and a frequency. The definition of the evoking strength is that "Given a patient with this finding, how strongly should I consider this diagnosis to be its explanation?" The frequency is an estimate of how often patients with the the disease have the finding. A scale of 0 to 5 and 1 to 5 in the INTERNIST-I knowledge base is represented as a shorthand version of evoking strengths and frequencies for judgmental information. For evoking strengths, the interpretation of "0" indicates nonspecific -- manifestation occurs too commonly to be used to construct a differential diagnosis. The interpretation of "5" indicates that the listed manifestation is pathognomonic for the diagnosis. The numbers from 1 to 4 interpret the importance between nonspecific and pathognomonic for the diagnosis. For frequencies, the interpretation of "1" is indicating that the listed manifestation occurs rarely in the disease. The interpretation of "5" indicates that the listed manifestation occurs in essentially all cases. The numbers from 2 to 4 interpret the importance of findings between rarely occurred and essentially occurred in all cases.

The score of each disease is calculated as the sum of a positive and a negative component. The positive component is based on the evoking strengths of the observed manifestations for the diagnosis. For example, an evoking strength of 0 counts as 1 point, a strength of 1 counts as 4 points, a 2 counts as 10 points, a 3 counts as 20, a 4 as 40 and a 5 as 80 points. The negative component includes the weight of all manifestations that are expected to occur in patients with the disease but are absent in the patient under consideration. The scale of the negative components is based on the expected frequency of the manifestation in the disease. For example, a frequency of 1 counts as -1 point, a 2 as -4 points, a 3 as -7 points, a 4 as -15 points, and a 5 as - 30 points.

The INTERNIST-I scoring system is more dependent on the knowledge engineer's (physician's) experience than other scoring systems which use statistical data. In INTERNIST-I knowledge base, the point count is nonlinear and the scale is arbitrary.

1.2 The HELP system

HELP (Health Evaluation through Logical Processing) is a Mainframe based Hospital Information System,[5] which contains expert system tools from which a system can be built to perform diagnosis. HELP is designed to meet educational, teaching as well as clinical needs at the Latter Day Saints (LDS) Hospital, a 550 bed teaching facility of the University of Utah School of Medicine. It has been underdeveloped for the past 16 years. This system is based on Tandem hardware consisting of 10 tandem central processor units and nearly 500 terminals as well as 80 printers in daily operation.

The HELP medical decision making system, which includes a frame-based, sequential Bayesian expert system and patient data base, was first implemented in 1974.[6] The HELP patient data base is generated from patient information sources throughout the LDS hospital. The patient data base includes patient history, some of physical exam, chemistry laboratory, hematologic laboratory, microbiology laboratory, ECG, radiology and some of specialized laboratory (such as pulmonary laboratory, cardiac catheterization laboratory and so on). The patient data base also provides information from the emergency room, pharmacy, nursing stations, admission, discharge, transition and others. The data collected and stored by the computer are used not only in managing patients but also in research, teaching and hospital management.

1.3 The Iliad medical expert system

An application of a computerized medical diagnosis system has been recently explored to assist in teaching third year medical students. To make such a system more accessible to students, the system is a microcomputer-based medical expert system called Iliad.[7]

Iliad is a frame based medical education system that runs on Macintosh computers. The Iliad system contains knowledge base (diagnostic frames, relation file and so on), knowledge base data dictionary and application programs as well as related literature. The Iliad knowledge base data dictionary (KBD) consists of an editor program and three documental files which are code, keywords and text file. A hierarchical structure allows the data to be accessed by general or specific terms. Figure 1 is an example of the KBD structure.

A user can search the medical terms by the keywords or code (in any hierarchical level) through the application programs. The diagnostic frames consist of probabilistic frames and deterministic frames (Bayesian and Boolean, detail in section 1.7).

The Iliad system performs two functions for medical students, teaching and testing. The teaching mode is Iliad's consulting configuration. In this mode, a student presents a real case to Iliad and Iliad generates a differential diagnosis. Various built-in teaching tools allow the student to inspect, study, and modify the consulting

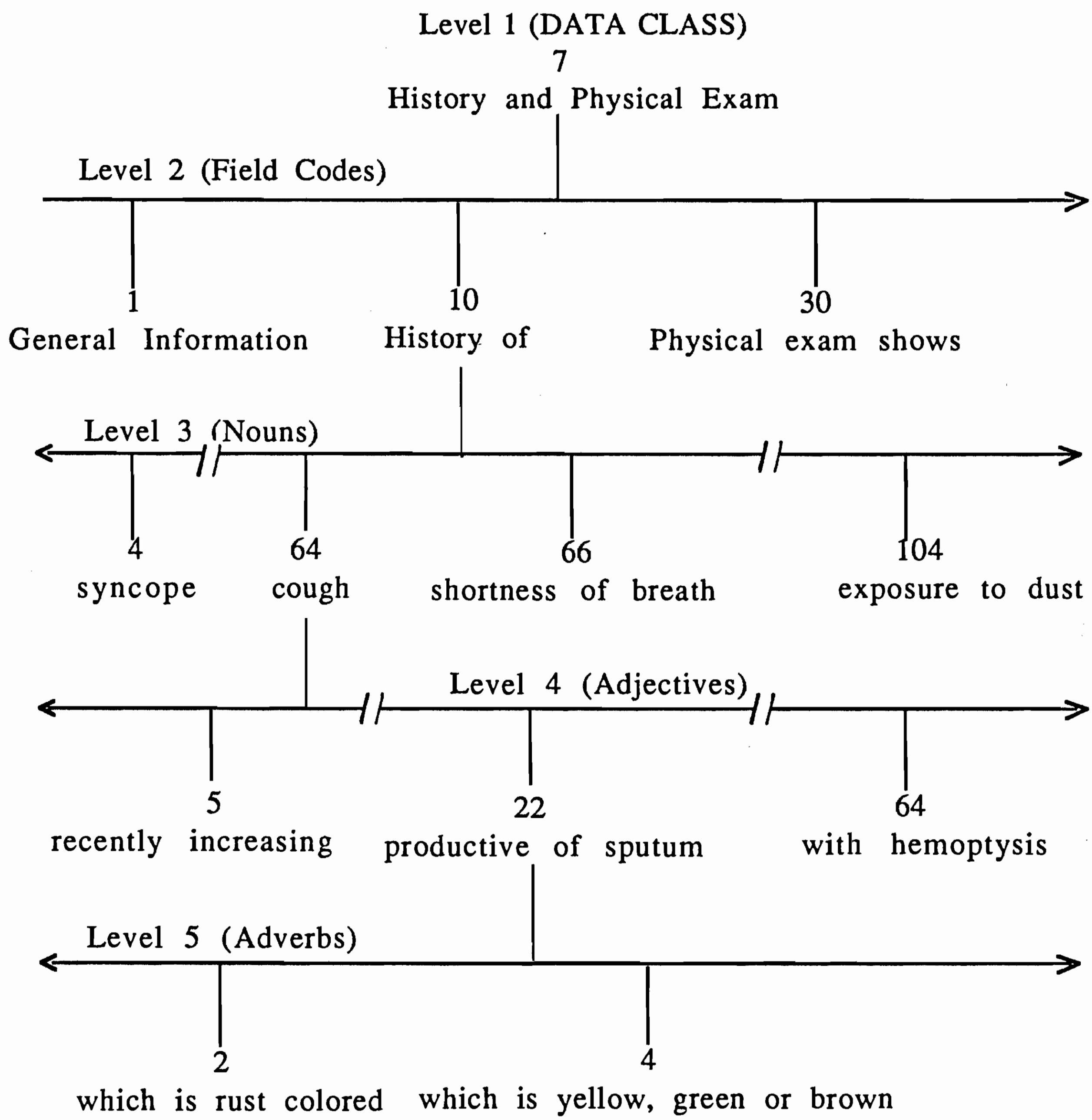

Figure 1. Hierarchical structure of the knowledge base data dictionary

diagnostic session. The testing mode is Iliad's simulation configuration. In the testing configuration, Iliad simulates an unknown, unique case. These cases can be generated anew from Iliad knowledge database each time. Iliad can select the simulation topic from the University's list of third-year medical student clerkship goals. Iliad presents the student with the chief complaint and allows the student to question the "patient." Iliad tracks the student's strategy and his diagnoses can be compared to what Iliad would have asked and concluded given the same information.

1.4 Importance of reliability in computer diagnosis

In developing the Iliad system, it became clear that accurate and reliable diagnostic probabilities were crucial. Students mistrusted the consultation mode when the probabilities in the differential diagnosis seemed unrealistically high or unrealistically low. In the simulator, accurate and reliable diagnoses are obviously essential if they are to be standards against which the student's performance is compared. Unfortunately, sequential Bayesian models rely on an assumption of conditional data independence. In medicine, many disease findings are interrelated, especially those that describe pathophysiologic concepts. For example, findings of fever, chills and white blood cell count increased in bacterial pneumonia are not independent: they tend to co-occur both in patients with the disease and without it. When the assumption of data independence is violated, sequential Bayesian systems become unreliable due to overconfidence. Iliad does not perform reliably under these conditions.

1.5 Compare different computerized expert systems

Several models of computer-aided prognosis have been compared by Ohmann, et al.,[8] which are Independence Bayes, Independence Bayes with global association factors, Independence Bayes combined with cluster analysis, Bayes with optimum tree dependence and linear logistic regression. In Ohmann's study, three points are helpful in understanding my research. The first point is about sample size, the larger the more accurate. Second, the performance of a model not only depends on the sample size but also on the number of variables (findings) used. There is no big difference among the results of these methods if few findings are used. The results would be different if many findings are used. The third point is that, with the same sample size and many findings, Independence Bayes combined with cluster analysis or Bayes with optimum tree dependence was markedly more accurate than Independence Bayes.

1.6 The solutions to solve conditionally dependent problem

In the model of independence Bayes, the conditional independence of the findings within each disease is assumed and knowledge frames rely only on sequential Bayesian calculations. Iliad's original knowledge base was this kind of independence Bayes. Because of conditional dependencies between patient findings in the Bayesian frames of this old knowledge base, Iliad was expected to produce overconfident decisions.[9] Realistic knowledge representation must take conditionally dependent, co-occurring findings into account. The old Iliad knowledge base attempted to account for conditionally dependent findings in one of two ways. The first solution was to use conditionally dependent findings as alternatives (using Boolean logic). The second solution was to eliminate most conditionally dependent findings, leaving only "key" findings (in the model of independence Bayesian). Unfortunately, these solutions were not completely successful.

The first solution restricted the diagnostic decision of the frame to either one conditionally dependent finding or another in any given patient. For example, there are five findings in the frame of asthma, which are asthma history, recurring pulmonary dysfunction, wheezing, asthma family history, prolonged expirations. These findings are conditionally dependent (co-occurring). If any three of these five findings present, the decision of the asthma frame is positive. Otherwise, the decision would be negative. This strategy is Boolean logic. One advantage of this approach was that conditionally dependent findings could remain in the frame. This strategy worked best when a small number of conditionally dependent findings represented diagnostic alternatives. The Boolean statements could arbitrate between the findings so that the most powerful finding was used in an individual patient case. For example, there are five findings in the frame of *Lung Consolidation*, which are cough, sputum, dyspnea, rales and radiographic lung infiltrate. Either the radiographic lung infiltrate is present or all of other four findings are present, the decision of the frame would be positive. Otherwise the decision would be negative. Unfortunately, this approach had three substantial disadvantages. First, separate Boolean logic statements had to be created for each frame, a time-consuming process. Second, these statements became increasingly complex when multiple conditional dependencies were represented. Third, each patient presents a different percentage of probability of the disease instead of simple "positive" or "negative."

The second solution is "sparse" frame. Sparse frames are different from Boolean frames, in that they eliminate all but "key" findings from the Bayesian frame. As in the first solution, *Lung Consolidation,* the sparse frame deleted the less specific physical findings of lung consolidation in favor of the "key" finding, radiographic lung infiltrate. Sparse frames can handle large numbers of conditionally dependent findings (simply eliminate them). However, this had significant educational disadvantages. Sparse frames were terse, noninformative to students exploring the knowledge base, and failed to respond appropriately when nonkey findings were entered during a patient consultation. Actual patients present with rich, diverse sets of findings, not "key" findings. The sparse frame model was simply unrealistic.

1.7 The concept of cluster

The solutions for conditional dependence provided by the old knowledge model were inadequate. Hence, a new model of knowledge representation, the clustered disease frame, has been developed. "Clusters" are groups of conditionally dependent disease findings that describe pathophysiologic states. Clusters generally use Boolean decision logic to return outcomes that can be used in Bayesian frames. The outcomes could be two to four categories, which can range from "confirmed" through "supported" and "suggested" to "denied." Figure 2 shows a typical cluster and illustrates these multiple potential outcomes.

These cluster outcomes can be passed to Bayesian frames. The Bayesian frames developed using the old knowledge model contained the conditionally dependent individual findings that are now encapsulated within clusters. The probabilities of cluster outcomes can be extracted from diseased and nondiseased populations (positive rate and negative rate). These new Bayesian frames revise the a priori disease prevalence according to cluster outcomes. This new type of Bayesian frame is illustrated in Figure 3.

1.8 Cluster vs noncluster

This project is to determine whether the clustered frame will result in more reliable diagnoses. In the method section, I will discuss a reliability study which is used to compare decision frames built using the old knowledge model with decision frames using the new, clustered model. A set of real patient data extracted from the HELP system patient database has been used to compare the new, clustered frames with the old, nonclustered frames. Because attempts had already been made to minimize overconfidence in the old frames (using Boolean or "key" findings), the procedure of comparing them to the new, clustered frames provides a conservative test of the hypothesized benefits of clustering. If the old frames had not been sparse, the benefits of clustering might have been even more pronounced.

Title: Chronic Airways Inflammation

Type: Boolean

Variables: prolonged cough **as** (having coughed daily for more then 2 months)
winter cough **as** (Cough daily during the winter months)
cough last year **as** (Having a similar cough a year age)
recurring cough **as** (Having spells of increased cough and sputum?)
morning cough **as** (Cough usually worse in the morning)
rhonchi **as** (rhonchi)

Logic: confirmed if [exist (prolonged cough) **or** exist (winter cough)] **and** [exist (cough last year) **or** exist (recurrent cough sputum)] then **true** else **false.**

suggested if exist (winter cough) **or** exist (prolonged cough) **or** exist (recurring cough sputum) **or** exist (morning cough) **or** exist (rhonchi) then **true** else **false.**

denied if not exist (prolonged cough) **and** not exist (recurring cough) **and** not exist (winter cough) **and** not exist (cough last year) **and** not exist (morning cough) **and** not exist (rhonchi) then **true** else **false.**

Figure 2. An example of the cluster *Chronic Airways Inflammation*

Title: Chronic Bronchitis diagnosis

Type: Sequential Bayesian

A priori: 0.0619

Cluster variables:	**confirmed**	**supported**	**suggested**	**denied**
Chronic airways inflammation				
(disease)	.44	---	.69	.31
(nondisease)	.14	---	.32	.71
Cigarette exposure				
(disease)	.63	---	---	.38
(nondisease)	.21	---	---	.79
generalized airway obstruction				
(disease)	.09	---	.31	.34
(nondisease)	.02	---	.23	.31

Nonclustered:	**TPR/FNR**	**FPR/TNR**
pulmonary toxin exposure	0.41/0.59	0.18/0.82
history of chronic bronchitis	0.38/0.62	0.03/0.97
xray COPD	0.59/0.41	0.06/0.94

Figure 3. An example of the *Chronic Bronchitis* Bayesian frame with clusters

The results section summarizes the reliability study for the frames of *Emphysema, Pneumothorax, Primary Neoplasm, Bacterial Pneumonia, Pulmonary Embolism, Chronic Bronchitis, Asthma, Metastases (METS) and Congestive Heart Failure (CHF).* It also gives insight into the ease of creating cluster-based disease frames. In many cases clinicians can achieve reliable clustering on the first try. However in other cases, the overconfidence can remain after a single attempt at clustering.

In the discussion section I discuss why more than one attempt at clustering is necessary and comment on the mistakes seen in the expert's estimation of sensitivity and specificity.

1.9 The hypotheses

My study examines three interrelated hypotheses.

First, substantially conditional dependence exists between findings in nonclustered Bayesian frames and this dependence leads to inaccuracies in assigning probabilities. These inaccuracies typically consist of a tendency to overestimate the probabilities of likely diagnoses and underestimate the probabilities of unlikely diagnoses. This behavior is referred to as "overconfidence."

Second, a clustered knowledge representation can reduce these inaccuracies by reducing the effects of conditional dependence in Bayesian frames.

Third, expert estimated statistics (true positive rate and false positive rate), influence, the cluster's outcomes are substantially different from real statistics.

If these hypotheses are proven, clusters provide a method of improving Iliad's diagnostic reliability. This improvement will facilitate the use of Iliad as a clinical consulting and teaching tool.

CHAPTER 2

METHOD

The description of the methods will be comprised of three parts. First, I will describe the approach to frame development using the clustered model. Second, I will describe the patient population within which the two knowledge models were compared. Third, I will describe the statistical procedures used to assess diagnostic reliability.

2.1 Frame development

The original knowledge frames were developed using the old, nonclustered knowledge representation model by Dr. Haug and other people as part of the HELP knowledge engineering project.[10] The new, clustered frames are direct descendents of these original frames, and contain exactly the same patient findings. The only difference is that the new frames contain clusters. These clusters were developed by Dr. Haug (a general internist), Dr. Lincoln (a pulmonary internist) and myself (a pediatrician). The structure and logic of these clusters are based on the experts' knowledge and experience. The clusters are frames that contain groups of conditionally dependent findings and often describe pathophysiologic concepts. Boolean logic is used within each cluster. The outcomes of cluster are multicategories which are "confirmed," "supported," "suggested" and "denied." For different clusters, the outcomes can be all of these four categories or only two or three of them. The new Bayesian frames have the same a priori prevalences as their old, nonclustered Bayesian counterparts. When individual findings are not clustered, these findings have the same sensitivity and specificity in both types of Bayesian frames. Pairs of Bayesian frames (old-new; nonclustered-clustered) were developed for the diseases of *chronic bronchitis, bacterial pneumonia, pulmonary embolism, asthma, pneumothorax, emphysema, metastasis, pulmonary neoplasm and congestive heart failure.*

2.2 The test population

The test population of this analysis is comprised of 517 patients who received chest radiographs while hospitalized in 1985 at LDS Hospital. I selected patients who had received a chest radiograph to ensure an adequate sample of patients in this test population with lung disease. Nevertheless, a proportion of the patients in the population had received incidental radiographs during an admission for nonpulmonary diseases. Each patient had a HELP database file that contained a medical history, some of his or her physical examinations, the radiographic result, and laboratory data gathered during hospitalization. The HELP system also contained the final ICD-9 diagnosis assigned to each patient. The ICD-9 code was used to indicate the "gold standard" diagnosis (i.e., the disease the patient really had).

2.3 The tools

I measured the a priori probability for each disease and true positive ratios (TPR, or sensitivity) and false positive ratios (FPR, or 1-specificity) for each individual finding as well as positive ratios, negative ratios for each cluster outcome in both diseased and nondiseased populations in the above patients' database. For the individual findings, the STRATO program was used to determine the TPR and FPR in the population of 517 patients. STRATO can stratify any patient database according to test characteristics. For instance, one can easily determine the conditional probability of hypoxemia given chronic bronchitis in a large patient database using STRATO. In statistical notation, this quantity is $P[Hypox^{+} | C.B.]$, or the sensitivity of hypoxemia in chronic bronchitis. For the newly-clustered frame items, I also required an accurate TPR and FPR for each cluster outcome. I wrote a set of programs in PTXT Application Language (PAL)[11] to stratify the patient database by cluster outcomes (confirmed, supported, suggested, and denied). These programs were run against the 517 patient database to derive the TPR an FPR for each outcome in a cluster.

Next, I converted the text format (shown in Figure 2) of the frames into PAL and incorporated the statistics derived from the patient data. The paired frames (nonclustered and clustered) are identical except for the clusters and their associated statistics. For instance, if the finding "fever" is a nonclustered finding, it appears identically in both the clustered and the nonclustered Bayesian diagnostic frames. The presence or absence of fever thus imparts the same diagnostic information in each case. Additionally, the a priori disease prevalences (also measured directly from the patient database) are identical between paired frames.

Once the paired diagnostic frames were coded in PAL, each frame was run against the patient database. For each of the 517 patients I determined the diagnostic probability (final posterior probability when all data were exhausted) of these nine diseases. These posterior likelihoods form the raw data for my reliability analysis.

These raw data are input to Microsoft Excel program that is a powerful integrated spreadsheet for the Macintosh. Excel provides fast, powerful calculating ability and can handle statistical problems. I did the statistical analysis in the Excel with the goal of testing my first two hypotheses (inaccuracy in nonclustered Bayesian frames and clustering can reduce the inaccuracy).

These measurements provided exact a priori disease prevalences and exact sensitivities and specificities for findings and cluster outcomes. One might argue that I should have derived these statistics in a separate population and then applied them to the test population. However, the primary concern is the effect of the clustered knowledge model on diagnostic reliability. I wished to eliminate any confounding variability among populations that might influence diagnostic reliability, such as inaccurate probability estimates. By providing perfect population statistics, I was able to provide the opportunity for each knowledge model to achieve its best diagnostic reliability, if only overconfidence did not occur.

Finally, I used the chi-square test to determine whether the experts' estimates deviated significantly from the true prevalences.

Squared errors were calculated for each pair of physician-database estimates.

2.4 Assessing reliability

Hilden, et al. have defined a series of statistics for assessing the reliability of probabilistic medical expert systems.[12,13,14] This reliability analysis can be described in two different ways. In the first case, the expert system provides a continuous estimate of the probability for each diagnostic decision. For example, the system provides an estimated probability for the presence of disease in a specific patient. In the second case, the expert system provides a dichotomous rather than a continuous estimate of the probability of the disease. That is, the expert system provides a decision that the disease is present or absent by choosing the disease with the highest likelihood. There is evidence that dichotomous probability assessments lose diagnostic information.[15] Nevertheless, I chose to assess reliability both ways because doctors are often forced to make dichotomous decisions.

The goal of Hilden's reliability assessment procedure is to determine whether the computer-based frames provide an overconfident, an underconfident (diffident), or an accurate (i.e., reliable) estimate of the rate of disease in the test population. Overconfidence refers to the tendency to assign probabilities too high to relatively likely diseases and probabilities too low to relatively unlikely diseases. Underconfidence refers to the opposite tendency, namely, the tendency to assign probabilities too low to relatively likely diseases and probabilities too high to relatively unlikely diseases. An overconfident physician would constantly conclude that his patients had specific diseases when there was in fact insufficient evidence. A diffident physician would continue to require additional testing after sufficient information was present for a reliable physician to conclude a diagnosis. Reliability is the ability to assign to the diseases in a differential diagnostic list probabilities consistent with the evidence available to support them.

In order to complete the reliability assessment, Hilden defines 10 statistics, arbitrarily denoted as Q1 through Q10. (See Table 1.)

Hilden divides these 10 statistics into two groups. Q1 through Q5 measure the diagnostic reliability of a probabilistic expert system over a continuous scale of diagnostic certainty from 0% to 100%. Q6 through Q10 are analogous to Q1 through Q5, but measure diagnostic reliability when the diagnosis is made in a dichotomy (deemed either present or absent) based on the probabilities associated with the individual diseases.

Q1 is the <u>actual</u> mean probability (summed over all patients in the test population) that the computer-based frame has assigned to the real diagnosis for each patient. In the present study, the real diagnosis is defined for each patient as the ICD-9 discharge code assigned by the medical staff at LDS hospital to the patients in our test population. Q2 is defined as the <u>expected</u> mean probability of the diagnosis made by the system. It is derived from the probabilities assigned to all of the diseases in all of the patients for whom the system has been run.

Table 1

Descution of Q1 through Q10

Definition	continuous	dichotomous
Actual average score	Q 1	Q 6
Expected average score	Q 2	Q 7
Reliability measure	Q3	Q8
Standard deviation	Q4	Q9
Test statistic*	Q5	Q10

* approximate 5% critical values is ± 1.96.

The difference (Q1 minus Q2) between actual and expected mean diagnostic probabilities is called Q3. The value of Q3 reflects the discrepancy between the computer's average estimate of the probability of the disease and the actual estimate of the probability of disease in the test population. If the expected mean value (Q2) based on the system's behavior over all of the possible diseases is higher than the actual mean value (Q1), then the computer is overconfident. Alternatively, if the expected mean is lower than the mean of the actual, then the system is diffident. Finally, if Q1 is not significantly different from Q2, then the system provides reliable estimates.

Apart from random fluctuations, Q3 averages zero for perfectly reliable systems. Q3 can be conceptualized as a statistic sampled from a normal distribution. Q3 can be converted into a standard score so that the value can be compared to a standard normal distribution. The statistic, Q4, is simply the standard deviation for the distribution of Q3. When Q3 is divided by Q4, the resulting value of Q5 can be treated as a standard score (or Z-score) from a standard normal distribution. If 95% of sample values of Q3 from a perfectly reliable system, the Q5 should be within absolute 1.96 (approximately 2 standard deviation units) from 0. If the absolute value of a sample of Q5 is greater than 1.96, then one must reject the null hypothesis that the computer produces reliable decisions.

Hilden gives an example demonstrating the interpretation of negative and positive values of Q3. Let us suppose the system sometimes unwarrantedly stakes a 100% certainty on pneumonia in certain patients. I will examine the effects of this decision on Q3 when the patient actually does or actually does not have the disease. Whether the patient has pneumonia or not, that patient's contribution to the Q2 score would always be $(1.0^2 + 0^2) = 1.0$. Now in the patient who <u>really has pneumonia</u>, that patient's contribution to Q1 would be 1.0. In this case the net Q3 is zero (1.0 - 1.0 = 0). The system has behaved reliably. However, in the patient <u>without pneumonia</u> the expert system was mistakenly overconfident in assigning a 100% certainty of disease. In this patient the Q1 contribution would be zero. Because Q2 is still 1.0 the net contribution to Q3 (Q1 - Q2) for the nondiseased patient is (0 - 1.0) = -1.0. This demonstrates that mistakenly overconfident systems tend to make Q3 negative. A similar analysis can be used to demonstrate that Q3 tends to be positive for underconfident (diffident) systems.

The Q6 through Q10 statistics are directly analogous to Q1 through Q5 except that the Q6 through Q10 statistics compare Non-Error Rate (NER). Each patient is assigned a discrete diagnosis (present/absent), rather than a probabilistic diagnosis (like P[chronic bronchitis] = 0.8). The diagnosis assigned is that with the highest probability calculated by the system.

Q6 is the actual Non-Error Rate (NER) which indicates the frequency with which the system has assigned the patient's real diagnosis. Q7 is the expected NER (assuming the null hypothesis of perfect reliability). Q8 is the difference between actual and expected NERs. Hilden has shown that Q8, like Q3, will be negative in overconfident systems, positive in diffident systems, and zero (apart from random fluctuations) in perfectly reliable systems. In similar fashion, Q9 is the standard deviation of Q8, and Q10 is the number of standard deviations Q8 varies from the mean. Hilden has demonstrated that if the absolute value of Q10 is greater than 1.96 (2 standard deviation units), one must reject the null hypothesis of system reliability. Like Q5, Q10 is also positive in underconfident systems and negative in overconfident systems.

This description makes it clear that Q5 and Q10 are the "key" statistics. They indicate both the direction of the unreliable tendency (positive or negative; underconfident or overconfident) and the magnitude of the unreliability. If Q5 and Q10 exceed an absolute value of 1.96, then the system is significantly unreliable. I hypothesized that the Q5 and Q10 statistics derived from the unclustered diagnostic frames would be significantly negative, denoting overconfidence. I also hypothesized that clustering these same frames would reduce overconfidence and bring the values for Q5 and Q10 within the limits of plus or minus 1.96.

Hilden's Q-test can be applied to more than two diseases at a time as long as the diseases are mutually exclusive and exhaustive. However, my analysis is limited to a 2 diseases system (diseased vs. nondiseased). If all of the posterior probabilities of these nine diseases add up to 1.0 (100%) then the Q-test can be used in the nine diseases system.

2.5 Mathematical formula

The detailed explanation of mathematical formulas are discussed as following:

If a sample population is large enough, the distribution of diagnostic probabilities (certainties) can be regarded as normal. Hilden defines several "Q statistics" to examine the reliability of these probability determinations. Statistics Q1 through Q5 compare the expert system's diagnosis with the patient's actual diagnosis, while Q6 through Q10 are Non-Error Rate (NER) measurements.

Q1 is the mean score for the probabilities assigned the patient's real disease by the expert system. If N is the total number of patients, and $P(d_i)$ is the expert system's estimate of the probability for the actual disease in each patient, then:

$$Q1 = 1/N * \sum P(d_i).$$

Q2 is the expected mean score for the probabilities assigned the patient's real disease by the expert system (the expected value of Q1). A perfect reliable expert system will produce a value for Q1 that is equal to Q2. If the probability that the i^{th} patient has the d^{th} disease is represented as Pid, then:

$$Q2 = \text{Expected}(Q1) = (1/N) * \Sigma\Sigma \text{Pid}^2$$

In an unreliable expert system, Q1 will deviate from Q2 in proportion to the degree of unreliability. This deviation is called Q3:

$$Q3 = Q1 - Q2$$

Note that overconfident systems will tend to make Q3 negative, as will systems that produce "wild guesses." Diffident (underconfident) systems will make Q3 positive.

From the system's output I am able to calculate a standard deviation for Q1 called Q4:

$$Q4^2 = \left(\frac{1}{N}\right)^2 \sum_i \left\{ \sum_d P_{id}[P_{id} - E(P_{id})^2] \right\}$$

The Q5 statistic is the ratio of Q3 to Q4. The Q5 is a z-statistic and should be normally distributed. If the absolute value of Q5 is less than 1.96, then the value of Q3 (the variation of actual from expected results) is less than two standard deviations from the mean. This implies that the null hypothesis of system reliability cannot be rejected at the 5% level.

The Non-Error Rate (NER) measures are Q6 through Q10. The "diagnostic" posterior probability is a value of 50%. Now, the NER statistics allow us to compare the system's diagnosis with the real discharge diagnosis. Q6 is the actual NER:

$$NER = Q6 = (1/N)\ \Sigma(\text{if matched then 1 otherwise 0}).$$

Q7 is the estimated NER:

$$E(NER) = Q7 = (1/N)\ \Sigma[\max P(id)].$$

Reliability is measured in terms of the difference between the actual and expected NERs. In a reliable system the difference, Q8, is small:

$$Q8 = Q6 - Q7$$

Q8 gives a rough idea of how far the actual NER is from the expected NER. The best diagnostic system would produce a result of Q8 = 0. The variance of Q8 is a binomial variance formula p(1-p), called Q9:

$$Q9^2 = (1/N)^2 \sum\{[\max P(id)][1-\max P(id)]\}.$$

An approximate standard normal test statistic is

$$Q10 = Q8/Q9$$

The two sets of reliability statistics, Q1 to Q5 and Q6 to Q10 or Q3 and Q8 are quite analogous. Both Q1 and Q6 are measures of the actual diagnostic probability. Both Q2 and Q7 are measures of the expected diagnostic probability. The number of standard deviations that the actual diagnostic probability varies from the expected probability is a measure of system reliability. Q5 and Q10 are measures of this deviation, expressed in confidence intervals. If the absolute values of the test statistics Q5 or Q10 are greater than 1.96 ($|Q5| > 1.96$ or $|Q10| > 1.96$) then the null hypothesis of system reliability is rejected at the significance level of 5%.

2.6 Using chi-square test

I used the chi-square (X^2) test to determine whether the experts' estimates deviated significantly from the true prevalences. Squared errors were calculated for each pair of physician-database estimates (estimated statistics vs real statistics).

CHAPTER 3

RESULTS

Table 2 through Table 4 summarize the results of the reliability analysis for the frames of *Emphysema, Pneumothorax, Primary Neoplasm, Bacterial Pneumonia, Pulmonary Embolism, Chronic Bronchitis, Asthma, Metastases (METS) and Congestive Heart Failure (CHF).* Table 2 is the summary of the frame results, for those frames which required only one attempt at clustering. Table 3 is the summary of the frame results, for frames requiring two attempts at clustering. Table 4 is the summary of the frame results, for frames requiring four attempts at clustering.

3.1 Information about the result of reliability analysis

In these three tables, the lines of Q5s are the summary statistic for the reliability analysis of the continuous (0 to 100%) diagnostic probabilities. The lines of Q10s are the summary statistic for the reliability analysis of Non-Error Rates using 0.5 as the threshold required to conclude a diagnosis. Because some information is always lost by classifying into two bins rather than by a continuous distribution,[16] the Q5 and Q10 reliability statistics sometimes diverge. For instance, the unclustered *asthma* frame looks reliable according to Q10 (+0.66) but the Q5 statistic (-2.16) indicates

Table 2
The statistics of five diseases with one attempt of clustering

	Bacteria Pneumonia		Pulmonary Embolis		Emphysema		Pneumothorax		Primary Neoplasm	
	Unclustered	Clustered	Unclustered	Clustered	Unclustered	Clustered	Unclustered	Clustered	Unclustered	Clustered
Q1	0.913	0.831	0.957	0.920	0.951	0.934	0.974	0.957	0.967	0.968
Q2	0.926	0.836	0.967	0.910	0.977	0.938	0.994	0.964	0.967	0.969
Q3	-0.013	-0.006	-0.010	0.010	-0.026	-0.004	-0.020	-0.007	0.0001	-0.001
Q4	0.005	0.007	0.004	0.006	0.004	0.005	0.001	0.005	0.004	0.004
Q5	-2.543*	-0.764	-2.533*	1.602	-7.231*	-0.688	-14.56*	-1.463	-0.015	-0.213
Q6	0.932	0.882	0.967	0.954	0.957	0.963	0.973	0.971	0.979	0.981
Q7	0.950	0.885	0.979	0.943	0.986	0.960	0.995	0.980	0.979	0.981
Q8	-0.017	-0.003	-0.012	0.011	-0.028	0.004	-0.023	-0.009	-0.001	-0.001
Q9	0.008	0.013	0.006	0.009	0.005	0.008	0.002	0.006	0.006	0.005
Q10	-2.066*	-0.258	-2.174*	1.161	-6.007*	0.483	-9.580*	-1.556	-0.103	-0.093

* Statistic indicates significant lack of reliability.

Table 3
The statistics of three diseases with two attempts of clustering

	Chronic Bronchitis			Asthma			METS		
	Unclustered	First Clustered	Second Clustered	Unclustered	First Clustered	Second Clustered	Unclustered	First Clustered	Second Clustered
Q1	0.881	0.912	0.920	0.938	0.924	0.889	0.946	0.944	0.948
Q2	0.934	0.927	0.930	0.947	0.938	0.896	0.942	0.933	0.939
Q3	-0.053	-0.015	-0.010	-0.009	-0.014	-0.008	0.004	0.010	0.010
Q4	0.005	0.006	0.006	0.004	0.005	0.006	0.005	0.006	0.006
Q5	-10.57*	-2.70*	-1.85	-2.16*	-2.59*	-1.21	0.771	1.746	1.729
Q6	0.901	0.936	0.942	0.965	0.946	0.919	0.967	0.977	0.977
Q7	0.951	0.950	0.954	0.960	0.959	0.928	0.962	0.958	0.962
Q8	-0.049	-0.014	-0.012	0.005	-0.013	-0.009	0.005	0.019	0.015
Q9	0.008	0.008	0.008	0.007	0.008	0.010	0.008	0.008	0.008
Q10	-6.17*	-1.67	-1.51	0.662	-1.725	-0.888	0.628	2.31	1.959

* Statistic indicates significant lack of reliability.

Table 4

The statistics of CHF with three attempts of clustering

	Congestive Heart Failure				
		First	Second	Third	Fourth
	Unclustered	Clustered	Clustered	Clustered	Clustered
Q1	0.880	0.896	0.867	0.866	0.865
Q2	0.920	0.955	0.895	0.883	0.876
Q3	-0.040	-0.059	-0.028	-0.017	-0.012
Q4	0.006	0.004	0.006	0.007	0.007
Q5	-7.193*	-13.96*	-4.608*	-2.549*	-1.700
Q6	0.896	0.899	0.888	0.890	0.901
Q7	0.943	0.969	0.925	0.919	0.912
Q8	-0.048	-0.070	-0.037	-0.029	-0.011
Q9	0.009	0.007	0.010	0.011	0.011
Q10	-5.393*	-10.57*	-3.671*	-2.710*	-0.994

* Statistic indicates significant lack of reliability.

significant unreliability. I did not assume reliable behavior unless the absolute values of both Q5 and Q10 were less than 1.96.

For each unclustered frame, Q5 showed statistically significant unreliability in the direction of overconfidence excepting *Primary Neoplasm and Metastases* (I will describe these two frames in detail in the discussion section). The Q10 statistic for unclustered Bayesian frames was in agreement with Q5 in each case with the exception of *Asthma*.

3.2 One attempt at clustering

In the cases of *Emphysema,* the unclustered Q5 is equal to -7.231 and Q10 is equal to -6.007. After the first attempt at clustering, the Q5 is equal to -0.688 and Q10 is equal to 0.483. In the case of *Pneumothorax,* the unclustered Q5 is -14.561 and Q10 is -9.58. After initial clustering, the Q5 is -1.463 and Q10 is -1.556. For *Primary Neoplasm,* both the unclustered and clustered frames proved reliability. In the case of *Pulmonary Embolism,* the unclustered Q5 is -2.533 and Q10 is -2.174. After first clustering, the Q5 is 1.602 and Q10 is 1.161. For *Bacteria Pneumonia,* the unclustered Q5 is -2.543 and Q10 is -2.066. After initial clustering, the Q5 equals -0.764 and Q10 equals -0.258. The initial attempt of clustering removed all statistically significant overconfidence. These statistics are shown in Table 2 under the disease names.

3.3 Two attempts at clustering

In the cases of *Chronic Bronchitis, Asthma and Metastases* the initial attempt at clustering did not produce reliable behavior and a second attempt was necessary. For the frame of Chronic Bronchitis, the Q5 of the unclustered frame is -10.57 and the Q10 of the unclustered frame is -6.17. After the first attempt at clustering, the Q5 became -2.7 and Q10 became -1.67. Although the Q10 was statistically significant, the Q5 did not reach the significant level after the first attempt at clustering. The second attempt at clustering produce reliable results. After the second clustering, the Q5 was equal to -1.85 and Q10 was equal to -1.51. They both were significantly reliable. In the case of *Asthma* and *Metastases*, the reliabilities actually deteriorated with the initial clustering. The Q5 of unclustered Asthma was equal to -2.16 and Q10 was equal to 0.662. The first attempt at an Asthma cluster produced an unreliable result with Q5 equal to -2.59 and Q10 equal to -1.725. For unclustered Metastases, Q5 equaled 0.771 and Q10 equaled 0.628. They both were significantly reliable. But the initial clustering made the Q5 equal 1.746 and Q10 equal 2.31. In the second attempt at clustering, the Q5s and Q10s of all three diseases were less than 1.96, which indicated significant reliability.

These results are shown in Table 3. Table 3 has three columns under each disease which are *Chronic Bronchitis, Asthma* and *Metastases.* The lefthand column is the result for the unclustered frames. The middle column is the result for the initial attempt of clustering, and the righthand column depicts the reclustered frames.

3.4 Four attempts at clustering

Clustering the diseases of *Congestive Heart Failure (CHF)* is more complex than the previous eight disease frames. Four attempts at clustering were done for the purpose of reducing overconfidence. For each attempt at clustering after the first, the absolute values of Q5 and Q10 were getting more and more close to the significant level which is 1.96, but the first three attempts at clustering never reached this level. Therefore, a fourth attempt was needed.

In the case of *CHF*, the unclustered Q5 was equal to -7.193 and Q10 was equal to -5.393. For the first attempt at clustering, Q5 was equal to -13.96 and Q10 was equal to -10.57. For the second clustering, Q5 was equal to -4.608 and Q10 was equal to -3.671. For the third clustering, Q5 was equal to -2.549 and Q10 was equal to -2.71. Neither the unclustered frame nor first to third attempts at clustering produced reliable results. Only the fourth attempt at clustering produced significant reliability. At that point, Q5 equaled -1.7 and Q10 equaled -0.994. Table 4 shows the statistical results of the unclustered frame and four attempts at clustering. There are five columns under the disease *Congestive Heart Failure* in Table 4. The leftmost column is the result for the unclustered frames. The middle three columns are the results of frames from first, second and third attempts at clustering and reclustering. And the righthand column depicts the result of the successful attempt at reclustered frames.

I will review the mistakes that led to these unreliable clustered frames and discovered mistakes in the discussion section. A second, third or fourth attempts at clustering corrected these mistakes and eliminated the overconfident tendencies.

3.5 Real statistics vs. expert's estimated statistics

In each case where a cluster result was used in a Bayesian frame the medical experts estimated TPRs and FPRs for each outcome. The expert's estimated statistics (TPRs and FPRs) were different from the statistics derived from HELP patient data file. Table 5 through Table 7 are the comparison of the expert's guessed statistics with the derived statistics for the two-bin to four-bin clusters. In each table, the first column is the cluster name, the middle columns are the expert's estimated statistics and the right most columns are the derived statistics. For each expert's estimated statistics column, there are two to four subcolumns corresponding to different clustering outcomes (these could be confirmed, supported, suggested or denied).

Table 8 displays the means of cluster statistics(TPRs and FPRs) within different groups of bins. In these nine diseases, there are eight clusters with two bins. Fifteen clusters give results divided into three bins and six clusters produce results distributed over four bins.

Table 9 and Table 10 show chi-square results which display the accuracy of physician estimated sensitivity (TPR) and 1-specificity (FPR). These results are deviated significantly from true values ($p = 0.005$).

Table 5
Two bins statistics

Statistics	Expert's guess		Real	
	positive	negative	positive	negative
Disease				
Incr.Lt. atrial pressure	.90	.10	.277	.723
Inadequate cardiac output	.50	.50	.815	.185
Paroxy.Noct.Dysp.	.60	.40	.338	.662
Orthopnea	.75	.25	.369	.631
Pulmonary toxin exposure	.10	.90	.406	.594
Emphysema	.70	.30	.455	.545
Pleural irritation	.30	.70	.345	.655
Nondisease				
Incr.Lt atrial pressure	.15	.85	.018	.976
Inadequate cardiac output	.10	.90	.531	.469
Paroxy.Noct.Dysp.	.05	.95	.137	.863
Orthopnea	.10	.90	.159	.841
Pulmonary toxin exposure	.02	.98	.177	.823
Emphysema (cluster)	.10	.90	.032	.968
Pleural irritation(pneumo)	.03	.97	.211	.789

Table 6
Three bins statistics

Statistics outcomes	Expert's guess conf.	sugg.	denied	Real conf.	sugg.	denied
Disease						
Hx of asthma	.60	.90	.10	.778	.667	.167
prior-Dx-CHF	.70	.80	.30	.431	.431	.292
Pulmon.Edema	.50	.90	.10	.400	.846	.123
Card.Decomens.	.60	.90	.10	.800	.031	.200
Pulmon.Inflammation	.70	.95	.05	.438	.688	.313
Cigar.Expos.(bronch.)	.80	.15	.05	.625	.375	0.0
Gener.Air.Obstr	.20	.75	.05	.094	.313	.344
Cigar.Expos.(emphys.)	.70	.80	.10	.625	.375	0.0
Constitutional. signs	.25	.50	.25	0.0	.800	.200
Signs of METS	.25	.50	.25	.133	.200	.667
Local.Airway.Obstruct.	.25	.25	.50	.267	.133	.667
Airspace disease	.98	.80	.02	.517	.897	.000
Pulmonary infection	.85	.95	.15	.793	.621	.241
Pleural irritation	.50	.70	.30	.067	.467	.533
Acute Respir^Disease	.70	.80	.20	.067	.467	.533
Nondisease						
Hx of asthma	.10	.20	.80	.074	.222	.733
prior-Dx-CHF	.10	.20	.90	.058	.169	.699
Pulmon.Edema	.05	.20	.80	.086	.653	.445
Card.Decomens.	.05	.10	.95	.146	.002	.845
Pulmon.Inflammation	.05	.15	.85	.136	.320	.711
Cigar.Expos.(bronch.)	.30	.10	.60	.206	.790	0.0
Gener.Air.Obstr	.02	.10	.80	.023	.227	.313
Cigar.Expos.(emphys.)	.20	.30	.80	.206	.790	0.0
Constitutional. signs	.05	.10	.85	.056	.295	.649
Signs of METS	.05	.05	.90	.008	.133	.861
Local.Airway.Obstruct.	.02	.05	.93	.098	.026	.881
Airspace disease	.01	.07	.99	.039	.568	.000
Pulmonary infection	.05	.10	.96	.170	.303	.035
Pleural irritation	.05	.10	.90	.104	.396	.603
Acute Respir^Disease	.05	.10	.90	.104	.396	.604

Table 7
Four bins statistics

Statistics	Expert's guess				Real		
	proba	supp	sugg	denied	proba	supp	denied
Disease							
atopy	.10	.30	.15	.45	.444	.778	.167
obstruction	.10	.75	.95	.05	.556	.722	.834
fluid retention	.70	.80	.60	.10	.031	.677	.816
pulm parench. loss	.60	.70	.80	.10	.727	.182	.546
lung cancer risks	.05	.75	.10	.10	0.0	.800	.267
loc-chest empty	.99	.30	.40	.01	.667	0.0	.333
Nondisease							
atopy	.02	.20	.15	.45	.102	.307	.650
obstruction	.05	.07	.15	.85	.038	.269	.601
fluid retention	.05	.15	.15	.80	.029	.321	.861
pulm parench. loss	.05	.10	.30	.90	.085	.055	.726
lung cancer risks	.005	.40	.05	.545	.010	.221	.795
loc-chest empty	.01	.02	.03	.99	.012	0.0	.004

Table 8

Mean of the differently clustered outcomes over all of the bins

Confidence level	Disease group Estimated	Disease group Real data	Nondisease group Estimated	Nondisease group Real data
Confirmed				
2 bin clusters	0.575	0.397	0.081	0.179
3 bin clusters	0.532	0.402	0.077	0.101
4 bin clusters	0.420	0.404	0.031	0.046
Supported				
3 bin clusters	0.710	0.487	0.128	0.353
4 bin clusters	0.600	0.527	0.157	0.196
Suggested				
4 bin clusters	0.500	0.347	0.138	0.196
Denied				
2 bin clusters	0.425	0.451	0.919	0.699
3 bin clusters	0.168	0.285	0.862	0.492
4 bin clusters	0.135	0.147	0.756	0.409

Table 9
Accuracy of physician estimates for "confirmed" bin according to number of findings per cluster

Disease group (Sensitivity estimates)

Number of clusters	Chi-squared results	df	significance
total 28 clusters	107.5	27	0.005
clusters with ≤ 4 findings	69.9	14	0.005
Clusters with ≥ 5 findings	37.7	12	0.005

Nondisease group (False positive ratio estimates)

Number of clusters	Chi-squared results	df	significance
total 28 clusters	825.1	27	0.005
Clusters with ≤ 4 findings	305.4	14	0.005
Clusters with ≥ 5 findings	519.7	12	0.005

Table 10
Accuracy of physician confidence estimates according to number of bins per cluster

Disease group (Sensitivity estimates):

Decision outcome	Chi-squared results	df	significance
Confirmed			
2 bin clusters	347.0	4	0.005
3 bin clusters	1106.8	15	0.005
4 bin clusters	81.9	3	0.005
Supported			
3 bin clusters	587.1	16	0.005
4 bin clusters	179.5	4	0.005
Suggested			
4 bin clusters	59.5	4	0.005
Denied			
2 bin clusters	12.1	3	0.025
3 bin clusters	266.8	13	0.005
4 bin clusters	8.8	2	0.025

Nondisease group (False positive rate estimates):

Decision outcome	Chi-squared results	df	significance
Confirmed			
2 bin clusters	171.9	4	0.005
3 bin clusters	66.1	15	0.005
4 bin clusters	8.4	3	0.025
Supported			
3 bin clusters	751.7	16	0.005
4 bin clusters	45.7	4	0.005
Suggested			
4 bin clusters	84.3	5	0.005
Denied			
2 bin clusters	6.4	3	0.025
3 bin clusters	178.6	13	0.005
4 bin clusters	125.0	4	0.025

CHAPTER 4

DISCUSSION

4.1 Preclustered and overclustered

The data clearly shows clustered frames can exhibit significantly less diagnostic overconfidence than corresponding nonclustered frames. In the cases of *Primary Neoplasm* and *Metastases*, some of the findings had already been grouped within the old, nonclustered frames. For example, the finding of "definite cancer history" consists of six individual findings that are "lung cancer, colon cancer, stomach cancer, kidney cancer, breast cancer and testicle cancer." The finding of "weight loss" consists of "losing weight and lost 10 pounds or more during this illness." The finding "pleuritic chest pain" consists of "chest pain increased by breathing deeply and chest pain increased by coughing." The finding of "multiple parenchymal masses" consists of "parenchymal mass, multiple nodules and multiple masses." Boolean logic had been used in these preclustered findings. Because the attempts had already been made to reduce the overconfidence through these groupings, the statistical results are significantly reliable before the final clustering.

There was concerned that "overclustering" might produce underconfident frames. In this experiment I only discovered one case of significant underconfidence (Q5 or Q10 > +1.96) in the initial attempt to cluster *Metastases*. Reliability analysis appears to be a powerful tool to detect both overconfident and underconfident frames.

4.2 Residual conditional dependence due to overconfidence

In five cases, *Emphysema, Pneumothorax, Primary Neoplasm, Pulmonary Embolus* and *Bacterial Pneumonia*, it was possible to achieve complete reliability after a single attempt of clustering. In contrast, *Chronic Bronchitis, Asthma* and *Metastases* exhibited overconfident after a single attempt of clustering. The overconfidence still remained after three attempts of clustering in the case of *Congestive Heart Failure*. In *Asthma* and the first attempt of *Congestive Heart Failure*, the overconfidence was actually worse. It was suspected that the *Chronic Bronchitis, Asthma, Metastases* and *Congestive Heart Failure* frames contained residual conditional dependence between findings that had not been detected on the initial attempt of clustering. In the frames of *Metastases, Asthma* and *Chronic Bronchitis*, a second attempt at clustering produced reliable behavior. However, in the case of *Congestive Heart Failure*, the first, second and third attempts at clustering did not produce reliable behavior, and the fourth attempt at clustering was necessary.

4.3 Reexamine the chronic bronchitis frame

Several errors were found after reexamination of the initial attemp to cluster *Chronic Bronchitis* frame. The most obvious error was the use of the finding "cough" in two places. "Cough" was used in both the *Chronic Cough* and the *Increased Airways Secretions* cluster. Because "cough" was in effect counted double in any coughing patient, the diagnosis of *Chronic Bronchitis* tended to be overconfident. In the reclustered frame, the *Chronic Cough* and the *Increased Airways Secretions* have been combined into a new cluster, *Signs of Airway Inflammation*, that only used cough once. Several smaller errors were also fixed. When the reclustered *Chronic Bronchitis* frame was tested, the value of Q5 had fallen to -1.85, indicating that the reclustering had produced acceptably reliable behavior. The example of the clustered and reclustered *Chronic Bronchitis* frame is shown in Figure 4.

4.4 Reexamine the asthma frame

Two different types of problems were found in the *Asthma* frame. The first problem was that the Boolean logic for producing a "denied" result in the cluster *Generalized Airways Obstruction* could never come true. Hence, every patient in the test population was diagnosed as having at least "suggested" *Generalized Airways Obstruction.* This was obviously a cause of unreliability. Two instances of the second type of problem had been found, which were failing to appropriately include conditionally dependent findings in clusters. In each case I simply placed these findings in the appropriate clusters, where they should have been in the first place. After these two types of problems were fixed, the Q5 value for the revised *Asthma* frame fell to -1.21, indicating reliable performance. The comparison of initial and second attempts of asthma clustering is shown in Figure 5.

Chronic Bronchitis

First attempt at cluster variables:

Chronic cough clustered findings:

Have you coughed daily for more then 2 months?
Do you cough daily during the winter months?
Did you have a similar cough a year age?
Do you have spells of increased cough and sputum?

Cigarette Exposure clustered findings:

Have you ever smoked cigarettes?
Have you smoked cigarettes for more than 10 years?

Increased Airway Secretions clustered findings:

Do you have spells of increased cough and sputum?
Is your cough usually worse in the morning?
physical examine: rhonchi

Generalized airway obstruction clustered findings:

physical examine: generalized wheezing
pulmonary function partial II
pulmonary function complete
pft airway obstruction **as** ([SCT] = airway obstruction)

Nonclustered findings:

Have you been exposed to large amounts of dust or fumes in the work place?
Do you have chronic bronchitis?
chest xray: emphysema/COPD

Figure 4. The initial attempt and second attempt of chronic bronchitis diagnosis frames

Second attempt of cluster variables:

<u>Chronic airways inflammation clustered findings:</u>

Have you coughed daily for more then 2 months?
Do you cough daily during the winter months?
Did you have a similar cough a year age?
Do you have spells of increased cough and sputum?
Is your cough usually worse in the morning?
rhonchi

<u>Cigarette exposure clustered findings:</u>

Have you ever smoked cigarettes?
Have you smoked cigarettes for more than 10 years?

<u>generalized airway obstruction clustered findings:</u>

physical examine: generalized wheezing
pulmonary function partial II
pulmonary function complete
pft airway obstruction as ([SCT] = airway obstruction

Nonclustered findings:

Have you been exposed to large amounts of dust or fumes in the work place?
Do you have chronic bronchitis?
chest xray: emphysema/COPD

Figure 4. Continued

Asthma

First attempt at cluster variables:

Atopy clustered findings:

allergic wheezing **as** (Do you wheeze due to an allergy?)

family history **as** (Do any of your blood relatives have allergies, eczema, or asthma?)

eczema history **as** (Have you ever had an eczema type rash?)

eosinophils **as** (eosin% * WBC)

Logic: If **exist** allergic wheezing or **exist** eczema history then status = **probable;**

*If **exist** family history then status = **suggested**;

*If **not exist** allergic wheezing and **not exist** eczema history and **not exist** family history and eosinophils **< 250** then status = **denied.**

Generalized airway obstruction clustered findings:

dyspnea **as** (Have you been short of breath with this illness?)

wheezing history **as** (Have you had wheezing with this illness?)

allergic wheezing as (Do you wheeze due to an allergy?)

infectious wheezing **as** (Do you wheeze due to an infection in your lungs?)

pulses paradoxicus **as** (Physical exam: pulses paradoxicus)

Figure 5. The comparison of initial attempt and second attempt of *Asthma* diagnosis frames

prolonged expirations **as** (Physical exam: prolonged expirations)

generalized wheezing **as** (Physical exam: generalized wheezing)

Aa^gradient **as** (Blood gas: 125 - pCO2 * 1.25 - pO2)

fiO2 **as** (Blood gas: %FIO2)

age

spirometry done **as** (Spirometry data)

pft airway obstruction **as** ([SCT] = AIRWAY OBSTRUCTION)

bronchodilator improved **as** (Spirogra = Improved Post Bronchodilator)

xray hyperinflation as (Hyperlucency/hyperinflation)

Logic: *If **exist** (spirometry done) and [**exist** (pft airway obstruction) or **exist** (bronchodilator improved)] then status = **confirmed**;

*If **exist** (generalized wheezing) or [**exist** (prolonged expirations) or **exist** (pulses paradoxicus)] and **exist** dyspnea then status = **likely**;

If **exist** wheezing history or **exist** allergic wheezing or **exist** infectious wheezing or **exist** xray hyperinflation then status = **supported**;

*If **exist** (generalized wheezing) and (Aa^gradient **>** **20**) and **not** **exist** (wheezing history) and **exist** (spirometry done) and **not** **exist** (pft airway obstruction) then status = **denied**.

Figure 5. Continued

Nonclustered findings:

asthma history **as** (Have you ever had asthma?)

*current asthma **as** (Are you having an asthma attack?)

recurring pulmonary dysfunction **as** (Do you frequently have tightness or stuffiness in your lungs? Does the lung discomfort come and go?)

Second attempt at cluster variables:

Atopy clustered findings:

allergic wheezing **as** (Do you wheeze due to an allergy?)

family history **as** (Do any of your blood relatives have allergies, eczema, or asthma?)

eczema history **as** (Have you ever had an eczema type rash?)

eosinophils **as** (eosin% * WBC)

Logic: If **exist** allergic wheezing or **exist** eczema history then status = **probable**;

*If **exist** family history or eosinophils >= 500 then status = **supported**;

*If **not exist** allergic wheezing and **not exist** eczema history and **not exist** family history and eosinophils < **500** then status = **denied**.

Generalized Airway Obstruction clustered findings:

*current asthma **as** (Are you having an asthma attack?)

dyspnea **as** (Have you been short of breath with this illness?)

wheezing history **as** (Have you had wheezing with this illness?)

allergic wheezing as (Do you wheeze due to an allergy?)

infectious wheezing **as** (Do you wheeze due to an infection in your lungs?)

pulses paradoxicus **as** (Physical exam: pulses paradoxicus)

Figure 5. Continued

prolonged expirations **as** (Physical exam: prolonged expirations)
generalized wheezing **as** (Physical exam: generalized wheezing)
Aa^gradient **as** (Blood gas: 125 - pCO2 * 1.25 - pO2)
fiO2 **as** (Blood gas: %FIO2)
age
spirometry done **as** (Spirometry data)
pft airway obstruction **as** ([SCT] = AIRWAY OBSTRUCTION)
bronchodilator improved **as** (Spirogra = Improved Post Bronchodilator)
xray hyperinflation as (Hyperlucency/hyperinflation)

Logic: *If **exist** (spirometry done) and [**exist** (pft airway obstruction) or **exist** (bronchodilator improved)] or **exist** current asthma then status = **confirmed**;

*If **exist** (current asthma) or (**exist** dyspnea and Aa^gradient > age/4) then status = **probable**;

If **exist** wheezing history or **exist** allergic wheezing or **exist** infectious wheezing or **exist** xray hyperinflation then status = **suggested**;

*If **not exist** (current asthma) and **not exist** (wheezing history) and **not exist** dyspnea and **not exist** (allergic wheezing) and **not exist** (xray hyperinflation) then status = **denied**.

Nonclustered findings:

asthma history **as** (Have you ever had asthma?)
recurring pulmonary dysfunction **as** (Do you frequently have tightness or stuffiness in your lungs? Does the lung discomfort come and go?)

* Different finding is used between two attempts of clustering.

Figure 5. Continued

4.5 Reexamine the CHF frame

The problem in clustering the frame of *Congestive Heart Failure* is complex. At the first attempt of clustering, the cluster did not appropriately include highly conditionally dependent findings. The Q5 and Q10 are much higher than the significance level. After the second attempt of clustering, two problems were still existed. These two problems reflected the same mistake. Some highly conditionally dependent findings were separated into two clusters. They should have been in one cluster. First, the clusters of *Right Sided Heart Failure* and *History of Heart Failure* were highly interdependent. Second, the clusters of *Radiographic Signs of Lung Edema* and *Signs of Inadequate Cardiac Output* were also highly conditionally dependent. In the third attempt at clustering, the clusters of *Radiographic Signs of Lung Edema* and *Signs of Inadequate Cardiac Output* were combined into a new cluster, *Signs of Lung Edema.* The clusters of *Right Sided Heart Failure* and *History of Heart Failure* were combined into *Right Sided Heart Failure.* The findings, "heart murmur history" and "tachycardia" are extracted from their previous cluster, *Right Sided Heart Failure* and *Signs of Lung Edema,* to be individual findings. The third attempt at clustering still did not produce a reliable result and one final mistake was found in this attempt to cluster. This mistake was based on the fact that the findings, "paroxysmal nocturnal dyspnea," "orthopnea," and "rales," were highly interdependent with some other findings which belonged to the cluster *Signs of Lung Edema.* After these problems were fixed, the Q5 value for the revised *Congestive Heart Failure* frame fell to -1.7 and Q10 fell to -0.99, indicating reliable performance. The details of the four attempts of *Congestive Heart Failure* clustering are shown in Figure 6.

4.6 Heuristic biases in experts' estimates

The representativeness heuristic causes a judge to evaluate the probability that object A belongs to a class B according to whether A resembles B. A physician who judges a patient's risk factors for pneumonia according to whether the patient appears coughing succumbs to representativeness. Availability is another powerful heuristic: a recent "string" of rare cases may cause one to overestimate the base rates of rare diseases. Anchoring occurs when a starting estimate is adjusted to arrive at the final answer. Adjustment from this starting point is typically inadequate and the final probability is in error.

The previous heuristic biases could prejudice both prevalence estimations of single findings and cluster outcomes. In fact, the cluster outcomes seem more complex. There has been little previous research on this heuristic.[17] They believe we construct mental models of possible events. For instance, a physician might construct a mental model of the prototypic patient's findings in a case of florid pulmonary edema. Decisions for situations (clusters) which closely resemble this model tend to be assigned high probabilities, while decisions for clusters which do not resemble the model are assigned low probabilities. This heuristic can lead to overly high probability assignments when a cluster decision seems quite typical of a disease, but is uncommon. On the other hand, a cluster decision that seems

Congestive Heart Failure

First attempt at clustered variables:

Fluid retention:
- Is your shortness of breath worse lying down than sitting up?
- Does your shortness of breath wake you up at night?
- Do you have to get up several times at night to urinate?
- Have you noticed swelling in your legs or ankles?
- pedal pitting edema

Previously Diagnosed Cardiac Failure:
- Have you had heart failure?
- Do you take a pill for your heart?
- Do you take digoxin, digitalis, or lanoxin?
- Are you taking medication for irregular heart beats?

Pulmonary edema:
- Have you been short of breath with this illness?
- exertional SOB as (Do you get short of breath with exertion?
- rales
- diffuse alveolar infiltrate on chest xray
- Kerly lines
- small irregular infiltrates on chest xray
- respiratory rate > 20/minuter

Increased lt atrial pressure:
- Increased jugular venous pressure
- Hepato-Jugular reflux
- Audible S3
- Pulmonary venous hypertension on chest xray

Cardiac decompensation:
- Audible S4
- PMI displaced laterally
- Cardio-Pericardio enlargement on chest xray

Nonclustered findings:

Have you been told you have a heart murmur?
Heart rate > 100/minute

Figure 6. Display of four attempts at clustering for disease of CHF

Second attempt at clustered variables:

Right sided heart failure present clustered findings:

Do you have to get up several times at night to urinate?
Have you noticed swelling in your legs or ankles?
Elevated jugular venous pressure
hepatojufular reflux
pedal pitting edema

Signs of inadequate cardiac output clustered findings:

Have you been short of breath with this illness?
Do you get short of breath with exertion?
Apical heart rate > 100/minute
Respiratory rate > 20/minute

Radiographic signs of lung edema:

diffuse alveolar infiltrate on chest xray
Kerly B lines
small irregular infiltrates on chest xray
diffuse alveolar infiltrate (fcm) Perihilar edema

Cardiac enlargement:

PMI displaced laterally
cardio-pericardial enlargement

History of heart failure:

heart failure history
arrhythmia history(take medication for irregular heart beat)
heart murmur history
inotropic drug use (take digoxin, digitalis or lanoxin)

Nonclustered findings:

Does your shortness of breath wake you at night?
Is your shortness of breath worse lying flat than sitting up?
rales (left basilar or right basilar rales)
S3 (Audible S3)
S4 (Audible S4)

Figure 6. Continued

Third attempt at clustered variables:

Right sided heart failure clustered findings:

Do you have to get up several times at night to urinate?
Have you noticed swelling in your legs or ankles?
Have you had heart failure?
Are you taking medication for irregular heart beats?
Do you take digoxin, digitalis, or lanoxin?
Elevated jugular venous pressure
hepatojufular reflux
pedal pitting edema

Signs of lung edema:

Have you been short of breath with this illness?
Do you get short of breath with exertion?
tachypnea (Respiratory rate > 20/minute)
diffuse alveolar infiltrate on chest xray
Kerly B lines
small irregular infiltrates on chest xray
perihilar edema [diffuse alveolar infiltrate (fcm) Perihilar edema]

Cardiac enlargement:

PMI displaced laterally
cardiopericardial enlargement on chest xray

Nonclustered findings:

Does your shortness of breath wake you at night?
Is your shortness of breath worse lying flat than sitting up?
Do you have a heart murmur?
rales (left basilar or right basilar rales)
S3 (Audible S3)
S4 (Audible S4)
tachycardia (Apical heart rate > 100/minuter)

Figure 6. Continued

Fourth attempt of clustered variables:

Right sided heart failure clustered findings:

Do you have to get up several times at night to urinate?
Have you noticed swelling in your legs or ankles?
Have you had heart failure?
Are you taking medication for irregular heart beats?
Do you take digoxin, digitalis, or lanoxin?
Elevated jugular venous pressure
hepatojufular reflux
pedal pitting edema

Signs of lung edema:

Does your shortness of breath wake you at night?
Is your shortness of breath worse lying flat than sitting up?
Have you been short of breath with this illness?
Do you get short of breath with exertion?
tachypnea (Respiratory rate > 20/minute)
rales (left basilar or right basilar rales)
diffuse alveolar infiltrate on chest xray
Kerly B lines
small irregular infiltrates on chest xray
perihilar edema

Cardiac enlargement:

PMI displaced laterally
cardiopericardial enlargement on chest xray

Nonclustered findings:

Do you have a heart murmur?
S3 (Audible S3)
S4 (Audible S4)
tachycardia (Apical heart rate > 100/minute)

Figure 6. Continued

nonspecific for the disease, but actually occurs quite frequently, may be assigned a mistakenly low probability.

4.7 Conclusion

Clusters are a new model of knowledge representation used in the Iliad expert system. Clusters encapsulate conditionally dependent findings and usually describe pathophysiologic entities. By reducing the effects of conditional dependence on Bayesian analysis, clusters increase the reliability of Iliad's decisions. Reliability analysis can discover unexpected overconfidence or underconfidence in newly developed knowledge frames. Ideally, knowledge frames should be debugged using reliability assessments before being used clinically. Unfortunately the lack of truly **comprehensive** clinical data bases precludes this.

The clustered knowledge model offers major advantages to a frame-based Bayesian decision system like Iliad. The most important advantage is that clusters increase diagnostic reliability. The data of reliability analysis clearly demonstrate that the clustered knowledge model is capable of increasing diagnostic reliability. Other advantages of clusters include modular knowledge representation, explicit teaching models for pattern recognition skills, and rich knowledge representations.

Clusters are powerful, hierarchical structures in which to encode medical knowledge. However, experts' estimation of cluster prevalences for Bayesian frames is difficult. Estimates of experts deviate significantly from actual prevalence values.

4.8 Future work

In this project, all of the clusterings are based on experts' experience. Sometimes clinicians do not estimate suitable clusters. Future work will provide accurate clusters which will be generated from real patient data by a statistical method. The clustered frames used in this study were based on sparse frames. Future work will also provide rich clustered frames. Of course, a reliability analysis to test rich, clustered frames against rich, unclustered frames is necessary.

REFERENCES

1 Gorry, G. A. and Barnett, G. O. Experience with a model of representation and utilization in a man-machine dialogue with a medical decision aid system. Methods Inf. Med. 21, 59 (1982).

2 Bayes, T. Essay towards solving a problem in the doctrine of chances. Phil. Trans. Roy. Soc. 53, 370 (1763); reprinted in Biometrika 45, 293 (1958).

3 Shortliffe, E. H., Davis, R., Axline, S. G., Buchannan, B. G., Green, C. C., and Cohen, S. N. Computer-based consultations in clinical therapertics: explanation and rule acquisition capabilities of the MYCIN system. Comput. Biomed. Res. 8, 303 (1975).

4 Miller, R. A., Pople, H. E., and Myers, J. D. INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. New Engl. J. Med. 307, 468 (1982).

5 Bouhaddou P, Haug PJ, Warner HR. Use of the HELP clinical database to build and test medical knowledge. Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care. IEEE Computer Society Press 1987:64-67.

6 Warner HR. Computer Assisted Decision Making. New York Academic Press 1979.

7 Hukill MJ, Ward KM, Haug PJ, Warner HR. HELP decision support on the Macintosh. Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care. IEEE Computer Society Press 1987:155-157.

8 Ohmann C, Yang Qin, Kunneke M, Stoltzing H, Thon K, Lorenz W. Bayes theorem and conditional dependence of symptoms: different models applied to data of upper gastrointestinal bleeding. Methods of Information in Medicine 1988; 27(2):73-83.

9 Weinstein MC, Fineberg, HV. Clinical decision analysis. 1st ed. W.B. Saunders Company 1980:156-158.

10 Haug P, Clayton PD, Shelton P, Rich T, Tocino I, Fredrick PR, Crapo RO, Morrison WJ. Revision of diagnostic logic using a clinical data base. Proceedings of AAMSI Congress 1987:238-242.

11 PTXT (Pointer to Text) is the HELP data dictionary

12 Habbema JDF, Hilden J, Bjeeregaard B. The measurement of performance in probabilistic diagnosis: The problem, descriptive tools, and measures based on classification matrices. Methodik der Information in der Medizin 1978; 17(4):217-226.

13 Hilden J, Habbema JDF, Bjerregaard B. The measurement of performance in probabilistic diagnosis: Trustworthiness of the exact values of the diagnostic probabilities. Methodik der Information in der Medizin 1978; 17(4):227-237.

14 Hilden J, Habbema JDF, Bjerregaard B. The measurement of performance in probabilistic diagnosis: Methods based on continuous function of the diagnostic probabilities. Methodik der Information in der Medizin 1978; 17(4):227-237.

15 Rifkin, RD. Maximum Shannon information content of diagnostic medical testing. Medical Decision Making 1985; 5(2):179-189.

16 Shapiro, AR. The evaluation of clinical predictions: A method and initial applications. New England Journal of Medicine 1977; 296:1509-1514.

17 Tversky A and Kahneman D. The simulation heuristic. In: Kahneman D, Slovic P, Tversky A, eds. Judgment under uncertainty: heuristics and biases. Cambridge: Cambridge University Press 1981:201-208.